

MÁSODIK GENERÁCIÓS SZEKVENÁLÁSI ELJÁRÁSOK
AZ AUJESZKY-FÉLE VÍRUS
TRANSZKRIPTÓMA-TÉRKÉPEZÉSÉBEN

SZTE Eötvös Loránd Kollégium

Az elmúlt évtized során a nagy léptékű genomikai vizsgálatok gyors fejlődése és költségeik csökkenése lehetővé tette számos élőlény teljes genetikai kódjának megismerését. Ezen technikák nélkülözhetetlenné váltak az élettudományokban, és az elkövetkező években rohamos fejlődésük mellett fokozatosan a napi laboratóriumi rutin részévé válnak, hasonlóan a nyolcvanas években széles körben elterjedt és számos módon adaptált polimeráz láncreakcióhoz. Segítségükkel a DNS mellett vizsgálható a hiszton-oktamerek DNS-hez való kötődése, a kromatin-kondenzáció, ezeknek különböző biokémiai módosulatai, valamint mindezek mellett az RNS-molekulák és azok különböző típusai, melyek az utóbbi évek eredményei alapján jóval fontosabb szerepet töltenek be az egyes sejtek és szervezetek közötti genetikai sokszínűség kialakításában, mint azt korábban feltételezték. A szöveti RNS-minták célzott szűrésével és cDNS-könyvtárak készítésével az adott szövet vagy szerv különböző RNS-típusai egymástól elkülönítve vizsgálhatók. Ezen vizsgálatok mutattak rá a mikroRNS-ek (miRNS) és hosszú nem kódoló RNS-ek (long non-coding RNA, lncRNA) jelentőségére – többek között az egyedfejlődésben [1] és a génkifejeződés szabályozásában [2]. A cDNS-szekvenálások elemzése emellett újszerű bioinformatikai megközelítést igényel, a molekulák sajátosságai miatt ugyanis alapvetően más jellegű kihívásokkal kell megküzdeni, mint a DNS-szekvenálás esetén.

2001-ben az Humán Genom Konzorcium és a Celera cég által párhuzamosan publikált első, közel teljes emberi genomszekvencia egy jelentős tudományos mérföldkövet jelentett. Mindkét, párhuzamosan futó genomprojekt az „első generációs”, azaz a lánctermináláson alapuló Sanger-féle szekvenálási eljárást alkalmazta a hárommilliárd bázispárnnyi szekvencia összeállításához. A Sanger-szekvenálás hátránya, hogy reakciónként maximum 800–1000 bázispárnnyi DNS-fragmens szekvenciája ismerhető meg, amely kísérleti megközelítéstől függően az egyes kromoszómák meghatározott szakaszairól származik (chromosome walking), illetve egy véletlenszerű lokuszról (shotgun sequencing). Míg az előző

esetben a megfelelően rendszerezett klónkönyvtárak elkészítése a leginkább munkaigényes feladat, az utóbbi módszer esetében a véletlenszerűen kapott DNS-szekvenciák összefüggő egészszé váló összefűzése jelenti a legnagyobb kihívást. Mindkét módszerről elmondható, hogy független laboratóriumok több mint egy évtizedes együttműködésével, több millió dolláros költségvetéssel valósultak meg.

Néhány évvel a Humán Genom Program sikeres lezárását követően azonban olyan technológiai ugrás következett be a genomikában, amelynek nyomán napjainkban ezt a rendkívüli tudományos teljesítményt jóval rövidebb idő alatt, töredéknyi költséggel képes reprodukálni egy arra specializálódott kutatócsoport. Az új, nagy áteresztőképességű készülékeket nevezzük második (ill. „új-”) generációs DNS-szekvenálóknak. Segítségükkel egy teljes emberi genom ma már ezer dollár körüli költséggel, egy-két hetes időkerettel feltérképezhető, > 99,9%-os pontossággal. Ezt a mikrofluidikában és a fluoreszcens jelölőmolekulák automatizált alkalmazásában elért fejlődés mellett az teszi lehetővé, hogy az elmúlt évtizedekben exponenciálisan növekvő számítási kapacitást kihasználva, hatékony algoritmusok segítségével az emberi genom egészen kisméretű, véletlenszerű fragmensekből is nagy pontossággal rekonstruálható. A mintaszámtól és a szekvenált mintamennyiségtől függően elérhető a teljes genomrekonstrukció, illetve a nagyobb, több száz kilobázispáros fragmensekből (kontigokból) álló „draft genome” összeállítása, amely később kiegészítő kísérletekkel véglegesíthető. Ennek a bioinformatikai ugrásnak az előfutára volt a Celera Genomics által kifejlesztett Celera Assembler [3], amelyet a Humán Genom Program idején használt technikáknak megfelelően még hosszabb fragmensekkel történő számításokra optimalizáltak. A DNS-szekvenálók második generációjának megjelenésével indultak fejlődésnek az olyan *de novo* genomillesztő algoritmusok, mint a SOAPdenovo vagy a Velvet [4]; illetve az olyan, memóriahatékony illesztőprogramok, amelyek új kísérletes adatokat egy, már meglévő referenciagenomra képesek illeszteni, ezzel megkerülve a rendkívül számításgényes *de novo* genomösszeillesztést a már ismert genommal rendelkező élőlények kutatásában. Az ilyen illesztőprogramok fő képviselői a Burrows-Wheeler Aligner ill. a Bowtie és annak cDNS-illesztésre optimalizált változata, a TopHat. Ezen algoritmusok, bár az elmúlt néhány évben számos továbbfejlesztett és bővített változatuk jelent meg, jelenleg is az összehasonlító genomikai kutatások alappilléreit jelentik.

A cDNS-szekvenálás során szintén érvényesül mind a térképezés, mind a *de novo* megközelítés, a DNS-szekvenálással ellentétben azonban az illesztés során nem nagyméretű kontigokat, hanem százezres nagyságrendű önálló transzkriptet kapunk, amelyek között számos eltérő izoforma fordul elő, amelyeket egyedi

splicehelyek és poszttranszlációs módosítások különböztetnek meg. Ezen izoformák elkülönítése jelenleg nem tökéletesen megoldott probléma. Emellett, míg DNS-szekvenálás során beállítható az a kívánt mintamennyiség, amellyel a teljes genom közel uniform módon lefedhető, Az RNS expresszió több nagyságrendnyi változást mutat az egyes transzkriptek között. Ezért lényeges a riboszomális RNS (rRNS) depléciónak megfelelő szintetikus komplementert tartalmazó elúciós oszlopokon, hiszen az rRNS tartalom általában 95% feletti. Depletálás mellett azonban szintén gyakori, hogy a teljes szekvenált cDNS-mennyiség jelentős százalékát néhány abundáns transzkript teszi ki, ezzel megakadályozva a ritkább RNS-formák detektálását. A fenti okok miatt a cDNS-szekvenálás körültekintő kísérlettervezést és annak megfelelő bioinformatikai elemzést kíván.

Az Aujeszký-féle vírus (Pseudorabies- PrV; *Suid herpesvirus 1*) az alphaherpeszvírusok családjába tartozó neurotrofpatogén, amely világszerte fertőző sertéspopulációkat, gyakran prenatális halálozást okozva [5]. Kedvező tulajdonságokkal rendelkezik mind a virológiai, mind a szélesebb körű transzkriptomikai kutatások szempontjából. Bár széles gazdaspektrumú vírus, embert nem fertőző, kompakt genomja pedig számos átfedő fehérjekódoló régiót tartalmaz, emellett jól elkülönülő gentetikai régiókra tagolható (unique long és unique short szakaszok, melyek közé két repeat-régió ékelődik). Mikrobák és vírusok esetén a génexpresszió vizsgálatát segíti, hogy nem kell számolni disztális enhanszer régiók hatásával. A PrV GC-tartalma extrém magas (72,4%), kódolt génjei a vírusfertőzés során időben és térben szigorúan szabályozott sorrendet követve, sok esetben kaszkádszerűen fejeződnek ki.

A vizsgálatok során a vírus propagációja PK-15 (porcine kidney-15) immortalizált sertésvese epitél sejtkultúrán történt, cDNS-szekvenáláshoz a vírusfertőzést követő 1, 2, 4, 6, 8, 12 órás időpontokból összeállított kevert totalRNS könyvtár készült. A minták szekvenálására Illumina és Pacific Biosciences platformokon került sor.

A bioinformatikai elemzés során kritikus lépés a vírus RNS-populációk elkülönítése a gazdaszervezetétől. Az extrém GC-tartalom, amellett, hogy figyelmet igényel a transzkriptabundanciák becslése során a szekvenáló platformok szisztematikus hibái miatt (GC-érzékenység a random hexamer amplifikáció következtében), segítséget jelent a többnyire átlagos GC-tartalmú eukarióta RNS-ektől és az esetleges kontaminációktól való elkülönítésben. A referenciagenomra történő megfelelő térképezést biztosítja a másodlagos adatok minőségi- és szekvenciahossz szerinti szűrése, valamint a TopHat algoritmus megfelelő paraméterezése, amely így nagy érzékenységgel detektálja a virális transzkripteken található alternatív splicehelyeket. Ezt követi az új, korábban még nem annotált transzkriptek

azonosítása, a gének abundancia szerinti rangsorolása, illetve az egynukleotidos polimorfizmusok, inszerciók és deléciók azonosítása és frekvenciájuk meghatározása. A kompakt genomszerveződés következtében számos antiszensz expressziót mutató régió azonosítható. Szerepük a génexpresszió szabályzásában többértű lehet, gátolhatják például a velük komplementer RNS-ekről történő fehérjeképződést, hatnak azonban direkt módon is a szomszédos génekre transzkripciós interferencián keresztül. Az antiszensz RNS-ek szerepe a virális génexpresszió szabályzásában jelenleg kevésbé kutatott terület, amelyben a második generációs szekvenálási technikák elterjedése jelentős előrelépést hozhat a korábban elérhetetlen felbontásnak és érzékenységnek köszönhetően.

IRODALOMJEGYZÉK

- [1] Guo F., Parker Kerrigan B. C., Yang D., Hu L., Shmulevich I., Sood A. K., Xue F, Zhang W.; *Journal of Hematology & Oncology* **7**:19, 2014
- [2] Garbett K. A., Vereczkei A., Kálmán S., Brown J. A., Taylor W. D., Faludi G., Korade Z., Shelton R.C., Mirmics K.; *Biological Psychiatry*, DOI: 10.1016/j.biopsych.2014.05.015, 2014
- [3] Myers E. W. *et al.*; *Science* **287**, 2196–2204., 2000
- [4] Zerbino D. R., Birney E.; *Genome Research* **18**, 821–829., 2008
- [5] Smith G.; *Annual Review of Microbiology* **66**, 153–176., 2012